B12

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C07H 21/04 — A1

(11) International Publication Number: WO 98/18967

(43) International Publication Date: 7 May 1998 (07.05.98)

(21) International Application Number: PCT/US97/19665

(22) International Filing Date: 27 October 1997 (27.10.97)

(30) Priority Data:

60/029,374	28 October 1996 (28.10.96)	US
08/813,508	7 March 1997 (07.03.97)	US

(71) Applicant *(for all designated States except US)*: AFFYMETRIX, INC. [US/US]; 3380 Central Expressway, Santa Clara, CA 95051 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: CHEE, Mark [AU/US]; 3199 Waverly Street, Palo Alto, CA 94306 (US). FAN, Jian-Bing [CN/US]; 275 Ventura Avenue #20, Palo Alto, CA 94306 (US).

(74) Agents: LIEBESCHUETZ, Joe et al.; Townsend and Townsend and Crew LLP, 8th floor, Two Embarcadero Center, San Francisco, CA 94111-3834 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: POLYMORPHISMS IN THE GLUCOSE-6 PHOSPHATE DEHYDROGENASE LOCUS

(57) Abstract

The invention provides nucleic acid segments of the glucose-6 phosphate dehydrogenase locus of the human genome including polymorphic sites. Allele-specific primers and probes hybridizing to regions flanking these sites are also provided. The nucleic acids, primers and probes are used in applications such as forensics, paternity testing, medicine and genetic analysis.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

POLYMORPHISMS IN THE GLUCOSE-6 PHOSPHATE DEHYDROGENASE LOCUS

CROSS-REFERENCE TO RELATED APPLICATION

The present application claims priority from provisional application 60/029,374, filed October 28, 1996, and USSN 08/813,508, filed March 7, 1997, which are incorporated by reference in their entirety for all purposes.

BACKGROUND OF THE INVENTION

The genomes of all organisms undergo spontaneous mutation in the course of their continuing evolution generating variant forms of progenitor sequences (Gusella, *Ann. Rev. Biochem.* 55, 831-854 (1986)). The variant form may confer an evolutionary advantage or disadvantage relative to a progenitor form or may be neutral. In some instances, a variant form confers a lethal disadvantage and is not transmitted to subsequent generations of the organism. In other instances, a variant form confers an evolutionary advantage to the species and is eventually incorporated into the DNA of many or most members of the species and effectively becomes the progenitor form. In many instances, both progenitor and variant form(s) survive and co-exist in a species population. The coexistence of multiple forms of a sequence gives rise to polymorphisms.

Several different types of polymorphism have been reported. A restriction fragment length polymorphism (RFLP) means a variation in DNA sequence that alters the length of a restriction fragment as described in Botstein et al., *Am. J.*

Hum. Genet. 32, 314-331 (1980). The restriction fragment length polymorphism may create or delete a restriction site, thus changing the length of the restriction fragment. RFLPs have been widely used in human and animal genetic analyses (see WO 90/13668; W090/11369; Donis-Keller, *Cell* 51, 319-337 (1987); Lander et al., *Genetics* 121, 85-99 (1989)). When a heritable trait can be linked to a particular RFLP, the presence of the RFLP in an individual can be used to predict the likelihood that the animal will also exhibit the trait.

Other polymorphisms take the form of short tandem repeats (STRs) that include tandem di-, tri- and tetra-nucleotide repeated motifs. These tandem repeats are also referred to as variable number tandem repeat (VNTR) polymorphisms. VNTRs have been used in identity and paternity analysis (US 5,075,217; Armour et al., *FEBS Lett.* 307, 113-115 (1992); Horn et al., W0 91/14003; Jeffreys, EP 370,719), and in a large number of genetic mapping studies.

Other polymorphisms take the form of single nucleotide variations between individuals of the same species. Such polymorphisms are far more frequent than RFLPs, STRs and VNTRs. Some single nucleotide polymorphisms occur in protein-coding sequences, in which case, one of the polymorphic forms may give rise to the expression of a defective or other variant protein and, potentially, a genetic disease. Examples of genes, in which polymorphisms within coding sequences give rise to genetic disease include β-globin (sickle cell anemia) and CFTR (cystic fibrosis). Other single nucleotide polymorphisms occur in noncoding regions. Some of these polymorphisms may also result in defective protein expression (e.g., as a result of defective splicing). Other single nucleotide polymorphisms have no phenotypic effects.

Single nucleotide polymorphisms can be used in the same manner as RFLPs, and VNTRs but offer several advantages. Single nucleotide polymorphisms occur with greater frequency and are spaced more uniformly throughout the genome than other forms of polymorphism. The greater frequency and uniformity of single nucleotide polymorphisms means that there is a greater probability that such a polymorphism will be found in close proximity to a genetic locus of interest than would be the case for other polymorphisms. Also, the different forms of characterized single nucleotide polymorphisms are often easier to distinguish that other types of polymorphism (e.g., by use of assays employing allele-specific hybridization probes or primers).

Despite the increased amount of nucleotide sequence data being generated in recent years, only a minute proportion of the total repository of polymorphisms in humans and other organisms has so far been identified. The paucity of polymorphisms hitherto identified is due to the large amount of work required for their detection by conventional methods. For example, a conventional approach to identifying polymorphisms might be to sequence the same stretch of oligonucleotides in a population of individuals by didoxy sequencing. In this type of approach, the amount of work increases in proportion to both the length of sequence and the number of individuals in a population and becomes impractical for large stretches of DNA or large numbers of persons.

SUMMARY OF THE INVENTION

The invention provides nucleic acid segments of between 10 and 100 bases containing at least 10, 15, 20, 25, 30 or 50 contiguous bases from any of the sequences shown in any of Tables 2-11 including a polymorphic site. Complements of these

segments are also included. The segments can be DNA or RNA, and can be double- or single-stranded. Some segments are 10-20 or 10-50 bases long. Preferred segments include a diallelic polymorphic site.

The invention further provides allele-specific oligonucleotides that hybridizes to a sequence shown in Tables 2-11 or its complement. These oligonucleotides can be probes or primers.

The invention further provides a method of analyzing a nucleic acid from an individual. The method determines which base is present at any one of the polymorphic sites shown in Tables 2-11. Optionally, a set of bases occupying a set of the polymorphic sites shown in Tables 2-11 is determined. This type of analysis can be performed on a plurality of individuals who are tested for the presence of a disease phenotype. The presence or absence of disease phenotype can then be correlated with a base or set of bases present at the polymorphic sites in the individuals tested.

DEFINITIONS

An oligonucleotide can be DNA or RNA, and single- or double-stranded. Oligonucleotides can be naturally occurring or synthetic, but are typically prepared by synthetic means. Preferred oligonucleotides of the invention include segments of DNA, or their complements including any one of the polymorphic sites shown in Tables 2-11. The segments are usually between 5 and 100 bases, and often between 5-10, 5-20, 10-20, 10-50, 20-50 or 20-100 bases. The polymorphic site can occur within any position of the segment. The segments can be from any of the allelic forms of DNA shown in Tables 2-11.

Hybridization probes are oligonucleotides capable of binding in a base-specific manner to a complementary strand of nucleic acid. Such probes include peptide nucleic acids, as described in Nielsen et al., *Science* 254, 1497-1500 (1991).

The term primer refers to a single-stranded oligonucleotide capable of acting as a point of initiation of template-directed DNA synthesis under appropriate conditions (*i.e.*, in the presence of four different nucleoside triphosphates and an agent for polymerization, such as, DNA or RNA polymerase or reverse transcriptase) in an appropriate buffer and at a suitable temperature. The appropriate length of a primer depends on the intended use of the primer but typically ranges from 15 to 30 nucleotides. Short primer molecules generally require cooler temperatures to form sufficiently stable hybrid complexes with the template. A primer need not reflect the exact sequence of the template but must be sufficiently complementary to hybridize with a template. The term primer site refers to the area of the target DNA to which a primer hybridizes. The term primer pair means a set of primers including a 5' upstream primer that hybridizes with the 5' end of the DNA sequence to be amplified and a 3', downstream primer that hybridizes with the complement of the 3' end of the sequence to be amplified.

Linkage describes the tendency of genes, alleles, loci or genetic markers to be inherited together as a result of their location on the same chromosome, and can be measured by percent recombination between the two genes, alleles, loci or genetic markers.

Polymorphism refers to the occurrence of two or more genetically determined alternative sequences or alleles in a population. A polymorphic marker or site is the locus at which divergence occurs. Preferred markers have at least two alleles,

each occurring at frequency of greater than 1%, and more preferably greater than 10% or 20% of a selected population. A polymorphic locus may be as small as one base pair. Polymorphic markers include restriction fragment length polymorphisms, variable number of tandem repeats (VNTR's), hypervariable regions, minisatellites, dinucleotide repeats, trinucleotide repeats, tetranucleotide repeats, simple sequence repeats, and insertion elements such as Alu. The first identified allelic form is arbitrarily designated as a the reference form and other allelic forms are designated as alternative or variant alleles. The allelic form occurring most frequently in a selected population is sometimes referred to as the wildtype form. Diploid organisms may be homozygous or heterozygous for allelic forms. A diallelic polymorphism has two forms. A triallelic polymorphism has three forms.

A single nucleotide polymorphism occurs at a polymorphic site occupied by a single nucleotide, which is the site of variation between allelic sequences. The site is usually preceded by and followed by highly conserved sequences of the allele (e.g., sequences that vary in less than 1/100 or 1/1000 members of the populations).

A single nucleotide polymorphism usually arises due to substitution of one nucleotide for another at the polymorphic site. A transition is the replacement of one purine by another purine or one pyrimidine by another pyrimidine. A transversion is the replacement of a purine by a pyrimidine or vice versa. Single nucleotide polymorphisms can also arise from a deletion of a nucleotide or an insertion of a nucleotide relative to a reference allele.

Hybridizations are usually performed under stringent conditions, for example, at a salt concentration of no more than 1 M and a temperature of at least 25°C. For example, conditions

of 5X SSPE (750 mM NaCl, 50 mM NaPhosphate, 5 mM EDTA, pH 7.4) and a temperature of 25-30°C are suitable for allele-specific probe hybridizations.

An isolated nucleic acid means an object species invention) that is the predominant species present (*i.e.*, on a molar basis it is more abundant than any other individual species in the composition). Preferably, an isolated nucleic acid comprises at least about 50, 80 or 90 percent (on a molar basis) of all macromolecular species present. Most preferably, the object species is purified to essential homogeneity (contaminant species cannot be detected in the composition by conventional detection methods).

DESCRIPTION OF THE PRESENT INVENTION

I. Novel Polymorphisms of the Invention

The human glucose-6-phosphate dehyrogenase locus (G6PD) encompasses more than 50,000 bp and resides on the X chromosome. A complete prototypical sequence of the G6PD locus has been published. That locus has remained relatively unexplored due to the cost and difficulty of conventional sequence analysis. The published sequence shows that the G6PD locus contains at least two genes, the G6PD gene and the 2-19 gene. Those genes span approximately 16,000 bp and 10,000 bp, respectively. The enzyme G6PD play a fundamental role in glucose metabolism. The function of the 2-19 polypeptide product, however, has not been shown.

The present application provides 10 polymorphisms at 10 nucleic acid sequence tagged sites in the human G6PD locus. Table 1 shows the base occupied at those ten sites in 10 individuals. The sequences flanking each of these polymorphic sites are shown in Tables 2-11. The base occupying the

polymorphic site is shown in bold in the table. In RNA forms of the claimed nucleic acids, a U replaces the T shown in Tables 2-11. The starting sequence and sequences designated M1-M10 are allelic variants. At each site analyzed, at last one of the M1-M10 is a novel allelic variant.

II. Analysis of Polymorphisms

A. Preparation of Samples

Polymorphisms are detected in a target nucleic acid from an individual being analyzed. For assay of genomic DNA, virtually any biological sample (other than pure red blood cells) is suitable. For example, convenient tissue samples include whole blood, semen, saliva, tears, urine, fecal material, sweat, buccal, skin and hair. For assay of cDNA or mRNA, the tissue sample must be obtained from an organ in which the target nucleic acid is expressed.

Many of the methods described below require amplification of DNA from target samples. This can be accomplished by e.g., PCR. *See generally PCR Technology: Principles and Applications for DNA Amplification* (ed. H.A. Erlich, Freeman Press, NY, NY, 1992); *PCR Protocols: A Guide to Methods and Applications* (eds. Innis, et al., Academic Press, San Diego, CA, 1990); Mattila et al., *Nucleic Acids Res.* 19, 4967 (1991); Eckert et al., *PCR Methods and Applications* 1, 17 (1991); *PCR* (eds. McPherson et al., IRL Press, Oxford); and U.S. Patent 4,683,202 (each of which is incorporated by reference for all purposes).

Other suitable amplification methods include the ligase chain reaction (LCR) (see Wu and Wallace, *Genomics* 4, 560 (1989), Landegren et al., *Science* 241, 1077 (1988), transcription amplification (Kwoh et al., *Proc. Natl. Acad. Sci. USA* 86, 1173 (1989)), and self-sustained sequence replication

(Guatelli et al., *Proc. Nat. Acad. Sci. USA*, 87, 1874 (1990)) and nucleic acid based sequence amplification (NASBA). The latter two amplification methods involve isothermal reactions based on isothermal transcription, which produce both single stranded RNA (ssRNA) and double stranded DNA (dsDNA) as the amplification products in a ratio of about 30 or 100 to 1, respectively.

B. Detection of Polymorphisms in Target DNA

There are two distinct types of analysis depending whether a polymorphism in question has already been characterized. The first type of analysis is sometimes referred to as de novo characterization. This analysis compares target sequences in different individuals to identify points of variation, i.e., polymorphic sites. By analyzing a groups of individuals representing the greatest ethnic diversity among humans and greatest breed and species variety in plants and animals, patterns characteristic of the most common alleles/haplotypes of the locus can be identified, and the frequencies of such populations in the population determined. Additional allelic frequencies can be determined for subpopulations characterized by criteria such as geography, race, or gender. The de novo identification of the polymorphisms of the invention is described in the Examples section. The second type of analysis is determining which form(s) of a characterized polymorphism are present in individuals under test. There are a variety of suitable procedures, which are discussed in turn.

1. Allele-Specific Probes

The design and use of allele-specific probes for analyzing polymorphisms is described by e.g., Saiki et al., *Nature* 324, 163-166 (1986); Dattagupta, EP 235,726, Saiki, WO 89/11548. Allele-specific probes can be designed that hybridize to a segment of target DNA from one individual but do not hybridize to the corresponding segment from another individual due to the presence of different polymorphic forms in the respective segments from the two individuals. Hybridization conditions should be sufficiently stringent that there is a significant difference in hybridization intensity between alleles, and preferably an essentially binary response, whereby a probe hybridizes to only one of the alleles. Some probes are designed to hybridize to a segment of target DNA such that the polymorphic site aligns with a central position (e.g., in a 15 mer at the 7 position; in a 16 mer, at either the 8 or 9 position) of the probe. This design of probe achieves good discrimination in hybridization between different allelic forms.

Allele-specific probes are often used in pairs, one member of a pair showing a perfect match to a reference form of a target sequence and the other member showing a perfect match to a variant form. Several pairs of probes can then be immobilized on the same support for simultaneous analysis of multiple polymorphisms within the same target sequence.

2. Tiling Arrays

The polymorphisms can also be identified by hybridization to nucleic acid arrays, some example of which are described by WO 95/11995 (incorporated by reference in its entirety for all purposes). One form of such arrays is described in the Examples section in connection with de novo identification of polymorphisms. The same array or a different

array can be used for analysis of characterized polymorphisms. WO 95/11995 also describes subarrays that are optimized for detection of a variant forms of a precharacterized polymorphism. Such a subarray contains probes designed to be complementary to a second reference sequence, which is an allelic variant of the first reference sequence. The second group of probes is designed by the same principles as described in the Examples except that the probes exhibit complementarily to the second reference sequence. The inclusion of a second group (or further groups) can be particular useful for analyzing short subsequences of the primary reference sequence in which multiple mutations are expected to occur within a short distance commensurate with the length of the probes (*i.e.*, two or more mutations within 9 to 21 bases).

3. Allele-Specific Primers

An allele-specific primer hybridizes to a site on target DNA overlapping a polymorphism and only primes amplification of an allelic form to which the primer exhibits perfect complementarily. See Gibbs, *Nucleic Acid Res.* 17, 2427-2448 (1989). This primer is used in conjunction with a second primer which hybridizes at a distal site. Amplification proceeds from the two primers leading to a detectable product signifying the particular allelic form is present. A control is usually performed with a second pair of primers, one of which shows a single base mismatch at the polymorphic site and the other of which exhibits perfect complementarily to a distal site. The single-base mismatch prevents amplification and no detectable product is formed. The method works best when the mismatch is included in the 3'-most position of the oligonucleotide aligned with the polymorphism because this position is most

destabilizing to elongation from the primer. *See, e.g.*, WO 93/22456.

4. Direct-Sequencing

The direct analysis of the sequence of polymorphisms of the present invention can be accomplished using either the dideoxy chain termination method or the Maxam Gilbert method (see Sambrook et al., *Molecular Cloning, A Laboratory Manual* (2nd Ed., CSHP, New York 1989); Zyskind et al., *Recombinant DNA Laboratory Manual*, (Acad. Press, 1988)).

5. Denaturing Gradient Gel Electrophoresis

Amplification products generated using the polymerase chain reaction can be analyzed by the use of denaturing gradient gel electrophoresis. Different alleles can be identified based on the different sequence-dependent melting properties and electrophoretic migration of DNA in solution. Erlich, ed., *PCR Technology, Principles and Applications for DNA Amplification*, (W.H. Freeman and Co, New York, 1992), Chapter 7.

6. Single-Strand Conformation Polymorphism Analysis

Alleles of target sequences can be differentiated using single-strand conformation polymorphism analysis, which identifies base differences by alteration in electrophoretic migration of single stranded PCR products, as described in Orita et al., *Proc. Nat. Acad. Sci.* 86, 2766-2770 (1989). Amplified PCR products can be generated as described above, and heated or otherwise denatured, to form single stranded amplification products. Single-stranded nucleic acids may refold or form secondary structures which are partially dependent on the base sequence. The different electrophoretic mobilities of single-

stranded amplification products can be related to base-sequence difference between alleles of target sequences.

III. Methods of Use

After determining polymorphic form(s) present in an individual at one or more polymorphic sites, this information can be used in a number of methods.

A. Forensics

Determination of which polymorphic forms occupy a set of polymorphic sites in an individual identifies a set of polymorphic forms that distinguishes the individual. *See generally* National Research Council, *The Evaluation of Forensic DNA Evidence* (Eds. Pollard et al., National Academy Press, DC, 1996). Since the polymorphic sites are within a 50,000 bp region in the human genome, the probability of recombination between these polymorphic sites is low. That low probability means the haplotype (the set of all 10 polymorphic sites) set forth in this application should be inherited without change for at least several generations. The more sites that are analyzed the lower the probability that the set of polymorphic forms in one individual is the same as that in an unrelated individual. Preferably, if multiple sites are analyzed, the sites are unlinked. Thus, polymorphisms of the invention are often used in conjunction with polymorphisms in distal genes. Preferred polymorphisms for use in forensics are diallelic because the population frequencies of two polymorphic forms can usually be determined with greater accuracy than those of multiple polymorphic forms at multi-allelic loci.

The capacity to identify a distinguishing or unique set of forensic markers in an individual is useful for forensic analysis. For example, one can determine whether a blood sample

from a suspect matches a blood or other tissue sample from a crime scene by determining whether the set of polymorphic forms occupying selected polymorphic sites is the same in the suspect and the sample. If the set of polymorphic markers does not match between a suspect and a sample, it can be concluded (barring experimental error) that the suspect was not the source of the sample. If the set of markers does match, one can conclude that the DNA from the suspect is consistent with that found at the crime scene. If frequencies of the polymorphic forms at the loci tested have been determined (e.g., by analysis of a suitable population of individuals), one can perform a statistical analysis to determine the probability that a match of suspect and crime scene sample would occur by chance.

p(ID) is the probability that two random individuals have the same polymorphic or allelic form at a given polymorphic site. In diallelic loci, four genotypes are possible: AA, AB, BA, and BB. If alleles A and B occur in a haploid genome of the organism with frequencies x and y, the probability of each genotype in a diploid organism are (see WO 95/12607):

Homozygote: $p(AA) = x^2$
Homozygote: $p(BB) = y^2 = (1-x)^2$
Single Heterozygote: $p(AB) = p(BA) = xy = x(1-x)$
Both Heterozygotes: $p(AB+BA) = 2xy = 2x(1-x)$

The probability of identity at one locus (i.e, the probability that two individuals, picked at random from a population will have identical polymorphic forms at a given locus) is given by the equation:

$$p(ID) = (x^2)^2 + (2xy)^2 + (y^2)^2.$$

These calculations can be extended for any number of polymorphic forms at a given locus. For example, the probability of identity p(ID) for a 3-allele system where the alleles have the frequencies in the population of x, y and z,

respectively, is equal to the sum of the squares of the genotype frequencies:

$p(ID) = x^4 + (2xy)^2 + (2yz)^2 + (2xz)^2 + z^4 + y^4$

In a locus of n alleles, the appropriate binomial expansion is used to calculate p(ID) and p(exc).

The cumulative probability of identity (cum p(ID)) for each of multiple unlinked loci is determined by multiplying the probabilities provided by each locus.

cum p(ID) = p(ID1)p(ID2)p(ID3).... p(IDn)

The cumulative probability of non-identity for n loci (i.e. the probability that two random individuals will be different at 1 or more loci) is given by the equation:

cum p(nonID) = 1-cum p(ID).

If several polymorphic loci are tested, the cumulative probability of non-identity for random individuals becomes very high (e.g., one billion to one). Such probabilities can be taken into account together with other evidence in determining the guilt or innocence of the suspect.

B. Paternity Testing

The object of paternity testing is usually to determine whether a male is the father of a child. In most cases, the mother of the child is known and thus, the mother's contribution to the child's genotype can be traced. Paternity testing investigates whether the part of the child's genotype not attributable to the mother is consistent with that of the putative father. Paternity testing can be performed by analyzing sets of polymorphisms in the putative father and the child.

If the set of polymorphisms in the child attributable to the father does not match the putative father, it can be concluded, barring experimental error, that the putative father

is not the real father. If the set of polymorphisms in the child attributable to the father does match the set of polymorphisms of the putative father, a statistical calculation can be performed to determine the probability of coincidental match.

The probability of parentage exclusion (representing the probability that a random male will have a polymorphic form at a given polymorphic site that makes him incompatible as the father) is given by the equation (see WO 95/12607):

$$p(exc) = xy(1-xy)$$

where x and y are the population frequencies of alleles A and B of a diallelic polymorphic site.

(At a triallelic site $p(exc) = xy(1-xy) + yz(1- yz) + xz(1-xz) + 3xyz(1-xyz))$, where x, y and z and the respective population frequencies of alleles A, B and C).

The probability of non-exclusion is

$$p(non\text{-}exc) = 1-p(exc)$$

The cumulative probability of non-exclusion (representing the value obtained when n loci are used) is thus:

$$cum\ p(non\text{-}exc) = p(non\text{-}exc1)p(non\text{-}exc2)p(non\text{-}exc3)\ldots. p(non\text{-}excn)$$

The cumulative probability of exclusion for n loci (representing the probability that a random male will be excluded)

$$cum\ p(exc) = 1 - cum\ p(non\text{-}exc).$$

If several polymorphic loci are included in the analysis, the cumulative probability of exclusion of a random male is very high. This probability can be taken into account in assessing the liability of a putative father whose polymorphic marker set matches the child's polymorphic marker set attributable to his/her father.

C. Correlation of Polymorphisms with Phenotypic Traits

The polymorphisms of the invention may contribute to the phenotype of an organism in different ways. Some polymorphisms occur within a protein coding sequence and contribute to phenotype by affecting protein structure. The effect may be neutral, beneficial or detrimental, or both beneficial and detrimental, depending on the circumstances. For example, a heterozygous sickle cell mutation confers resistance to malaria, but a homozygous sickle cell mutation is usually lethal. Other polymorphisms occur in noncoding regions but may exert phenotypic effects indirectly via influence on replication, transcription, and translation. A single polymorphism may affect more than one phenotypic trait. Likewise, a single phenotypic trait may be affected by polymorphisms in different genes. Further, some polymorphisms predispose an individual to a distinct mutation that is causally related to a certain phenotype.

Phenotypic traits include diseases that have known but hitherto unmapped genetic components. Phenotypic traits also include symptoms of, or susceptibility to, multifactorial diseases of which a component is or may be genetic, such as autoimmune diseases, inflammation, cancer, diseases of the nervous system, and infection by pathogenic microorganisms. Some examples of autoimmune diseases include rheumatoid arthritis, multiple sclerosis, diabetes (insulin-dependent and non-independent), systemic lupus erythematosus and Graves disease. Some examples of cancers include cancers of the bladder, brain, breast, colon, esophagus, kidney, leukemia, liver, lung, oral cavity, ovary, pancreas, prostate, skin, stomach and uterus. Phenotypic traits also include characteristics such as longevity, appearance (e.g., baldness, obesity), strength, speed, endurance, fertility, and

susceptibility or receptivity to particular drugs or therapeutic treatments.

Correlation is performed for a population of individuals who have been tested for the presence or absence of a phenotypic trait of interest and for polymorphic markers sets. To perform such analysis, the presence or absence of a set of polymorphisms (i.e. a polymorphic set) is determined for a set of the individuals, some of whom exhibit a particular trait, and some of which exhibit lack of the trait. The alleles of each polymorphism of the set are then reviewed to determine whether the presence or absence of a particular allele is associated with the trait of interest. Correlation can be performed by standard statistical methods such as a K-squared test and statistically significant correlations between polymorphic form(s) and phenotypic characteristics are noted. For example, it might be found that the presence of allele A1 at polymorphism A correlates with heart disease. As a further example, it might be found that the combined presence of allele A1 at polymorphism A and allele B1 at polymorphism B correlates with increased milk production of a farm animal.

Such correlations can be exploited in several ways. In the case of a strong correlation between a set of one or more polymorphic forms and a disease for which treatment is available, detection of the polymorphic form set in a human or animal patient may justify immediate administration of treatment, or at least the institution of regular monitoring of the patient. Detection of a polymorphic form correlated with serious disease in a couple contemplating a family may also be valuable to the couple in their reproductive decisions. For example, the female partner might elect to undergo in vitro fertilization to avoid the possibility of transmitting such a polymorphism from her husband to her offspring. In the case of

a weaker, but still statistically significant correlation between a polymorphic set and human disease, immediate therapeutic intervention or monitoring may not be justified. Nevertheless, the patient can be motivated to begin simple life-style changes (e.g., diet, exercise) that can be accomplished at little cost to the patient but confer potential benefits in reducing the risk of conditions to which the patient may have increased susceptibility by virtue of variant alleles. Identification of a polymorphic set in a patient correlated with enhanced receptiveness to one of several treatment regimes for a disease indicates that this treatment regime should be followed.

For animals and plants, correlations between characteristics and phenotype are useful for breeding for desired characteristics. For example, Beitz et al., US 5,292,639 discuss use of bovine mitochondrial polymorphisms in a breeding program to improve milk production in cows. To evaluate the effect of mtDNA D-loop sequence polymorphism on milk production, each cow was assigned a value of 1 if variant or 0 if wildtype with respect to a prototypical mitochondrial DNA sequence at each of 17 locations considered. Each production trait was analyzed individually with the following animal model:

$$Y_{ijkpn} = \mu + YS_i + P_j + X_k + \beta_1 + \ldots \beta_{17} + PE_n + a_n + e_p$$

where Y_{ijknp} is the milk, fat, fat percentage, SNF, SNF percentage, energy concentration, or lactation energy record; μ is an overall mean; YS_i is the effect common to all cows calving in year-season; X_k is the effect common to cows in either the high or average selection line; β_1 to β_{17} are the binomial regressions of production record on mtDNA D-loop sequence polymorphisms; PE_n is permanent environmental effect common to all records of cow n; a_n is effect of animal n and is composed

of the additive genetic contribution of sire and dam breeding values and a Mendelian sampling effect; and e_p is a random residual. It was found that eleven of seventeen polymorphisms tested influenced at least one production trait. Bovines having the best polymorphic forms for milk production at these eleven loci are used as parents for breeding the next generation of the herd.

D. Genetic Mapping of Phenotypic Traits

The previous section concerns identifying correlations between phenotypic traits and polymorphisms that directly or indirectly contribute to those traits. The present section describes identification of a physical linkage between a genetic locus associated with a trait of interest and polymorphic markers that are not associated with the trait, but are in physical proximity with the genetic locus responsible for the trait and co-segregate with it. Such analysis is useful for mapping a genetic locus associated with a phenotypic trait to a chromosomal position, and thereby cloning gene(s) responsible for the trait. *See* Lander et al., *Proc. Natl. Acad. Sci. (USA)* 83, 7353-7357 (1986); Lander et al., *Proc. Natl. Acad. Sci. (USA)* 84, 2363-2367 (1987); Donis-Keller et al., *Cell* 51, 319-337 (1987); Lander et al., *Genetics* 121, 185-199 (1989)). Genes localized by linkage can be cloned by a process known as directional cloning. *See* Wainwright, *Med. J. Australia* 159, 170-174 (1993); Collins, *Nature Genetics* 1, 3-6 (1992) (each of which is incorporated by reference in its entirety for all purposes).

Linkage studies are typically performed on members of a family. Available members of the family are characterized for the presence or absence of a phenotypic trait and for a set of polymorphic markers. The distribution of polymorphic markers in an informative meiosis is then analyzed to determine which

polymorphic markers co-segregate with a phenotypic trait. *See, e.g.*, Kerem et al., *Science* 245, 1073-1080 (1989); Monaco et al., *Nature* 316, 842 (1985); Yamoka et al., *Neurology* 40, 222-226 (1990); Rossiter et al., *FASEB Journal* 5, 21-27 (1991).

Linkage is analyzed by calculation of LOD (log of the odds) values. A lod value is the relative likelihood of obtaining observed segregation data for a marker and a genetic locus when the two are located at a recombination fraction θ, versus the situation in which the two are not linked, and thus segregating independently (Thompson & Thompson, *Genetics in Medicine* (5th ed, W.B. Saunders Company, Philadelphia, 1991); Strachan, "Mapping the human genome" in *The Human Genome* (BIOS Scientific Publishers Ltd, Oxford), Chapter 4). A series of likelihood ratios are calculated at various recombination fractions (θ), ranging from θ = 0.0 (coincident loci) to θ = 0.50 (unlinked). Thus, the likelihood at a given value of θ is: probability of data if loci linked at θ to probability of data if loci unlinked. The computed likelihoods are usually expressed as the $\log_{10}$ of this ratio (i.e., a lod score). For example, a lod score of 3 indicates 1000:1 odds against an apparent observed linkage being a coincidence. The use of logarithms allows data collected from different families to be combined by simple addition. Computer programs are available for the calculation of lod scores for differing values of θ (e.g., LIPED, MLINK (Lathrop, *Proc. Nat. Acad. Sci. (USA)* 81, 3443-3446 (1984)). For any particular lod score, a recombination fraction may be determined from mathematical tables. *See* Smith et al., *Mathematical tables for research workers in human genetics* (Churchill, London, 1961); Smith, *Ann. Hum. Genet.* 32, 127-150 (1968). The value of θ at which the lod score is the highest is considered to be the best estimate of the recombination fraction.

Positive lod score values suggest that the two loci are linked, whereas negative values suggest that linkage is less likely (at that value of θ) than the possibility that the two loci are unlinked. By convention, a combined lod score of +3 or greater (equivalent to greater than 1000:1 odds in favor of linkage) is considered definitive evidence that two loci are linked. Similarly, by convention, a negative lod score of -2 or less is taken as definitive evidence against linkage of the two loci being compared. Negative linkage data are useful in excluding a chromosome or a segment thereof from consideration. The search focuses on the remaining non-excluded chromosomal locations.

IV. Modified Polypeptides and Gene Sequences

The invention further provides variant forms of nucleic acids and corresponding proteins. The nucleic acids comprise at least ten contiguous bases of one of the sequences described in Tables 2-11 designated M1-M10. Some nucleic acid encode full-length variant forms of proteins. Similarly, variant proteins have the prototypical amino acid sequences of encoded by nucleic acid sequence shown in Tables 2-11, designated M1-M10 (read so as to be in-frame with the full-length coding sequence of which it is a component).

Variant genes can be expressed in an expression vector in which a variant gene is operably linked to a native or other promoter. Usually, the promoter is a eukaryotic promoter for expression in a mammalian cell. The transcription regulation sequences typically include a heterologous promoter and optionally an enhancer which is recognized by the host. The selection of an appropriate promoter, for example trp, lac, phage promoters, glycolytic enzyme promoters and tRNA promoters, depends on the host selected. Commercially available expression

vectors can be used. Vectors can include host-recognized replication systems, amplifiable genes, selectable markers, host sequences useful for insertion into the host genome, and the like.

The means of introducing the expression construct into a host cell varies depending upon the particular construction and the target host. Suitable means include fusion, conjugation, transfection, transduction, electroporation or injection, as described in Sambrook, *supra*. A wide variety of host cells can be employed for expression of the variant gene, both prokaryotic and eukaryotic. Suitable host cells include bacteria such as *E. coli*, yeast, filamentous fungi, insect cells, mammalian cells, typically immortalized, *e.g.*, mouse, CHO, human and monkey cell lines and derivatives thereof. Preferred host cells are able to process the variant gene product to produce an appropriate mature polypeptide. Processing includes glycosylation, ubiquitination, disulfide bond formation, general post-translational modification, and the like.

The protein may be isolated by conventional means of protein biochemistry and purification to obtain a substantially pure product, *i.e.*, 80, 95 or 99% free of cell component contaminants, as described in Jacoby, *Methods in Enzymology* Volume 104, Academic Press, New York (1984); Scopes, *Protein Purification, Principles and Practice*, 2nd Edition, Springer-Verlag, New York (1987); and Deutscher (ed), *Guide to Protein Purification, Methods in Enzymology*, Vol. 182 (1990). If the protein is secreted, it can be isolated from the supernatant in which the host cell is grown. If not secreted, the protein can be isolated from a lysate of the host cells.

The invention further provides transgenic nonhuman animals capable of expressing an exogenous variant gene and/or

having one or both alleles of an endogenous variant gene inactivated. Expression of an exogenous variant gene is usually achieved by operably linking the gene to a promoter and optionally an enhancer, and microinjecting the construct into a zygote. *See* Hogan et al., "Manipulating the Mouse Embryo, A Laboratory Manual," Cold Spring Harbor Laboratory. Inactivation of endogenous variant genes can be achieved by forming a transgene in which a cloned variant gene is inactivated by insertion of a positive selection marker. *See* Capecchi, *Science* 244, 1288-1292 (1989). The transgene is then introduced into an embryonic stem cell, where it undergoes homologous recombination with an endogenous variant gene. Mice and other rodents are preferred animals. Such animals provide useful drug screening systems.

In addition to substantially full-length polypeptides expressed by variant genes, the present invention includes biologically active fragments of the polypeptides, or analogs thereof, including organic molecules which simulate the interactions of the peptides. Biologically active fragments include any portion of the full-length polypeptide which confers a biological function on the variant gene product, including ligand binding, and antibody binding. Ligand binding includes binding by nucleic acids, proteins or polypeptides, small biologically active molecules, or large cellular structures.

Polyclonal and/or monoclonal antibodies that specifically bind to variant gene products but not to corresponding prototypical gene products are also provided. Antibodies can be made by injecting mice or other animals with the variant gene product or synthetic peptide fragments thereof. Monoclonal antibodies are screened as are described, for example, in Harlow & Lane, *Antibodies, A Laboratory Manual*, Cold Spring Harbor Press, New York (1988); Goding, *Monoclonal*

antibodies, Principles and Practice (2d ed.) Academic Press, New York (1986). Monoclonal antibodies are tested for specific immunoreactivity with a variant gene product and lack of immunoreactivity to the corresponding prototypical gene product. These antibodies are useful in diagnostic assays for detection of the variant form, or as an active ingredient in a pharmaceutical composition.

V. Kits

The invention further provides kits comprising at least one allele-specific oligonucleotide as described above. Often, the kits contain one or more pairs of allele-specific oligonucleotides hybridizing to different forms of a polymorphism. In some kits, the allele-specific oligonucleotides are provided immobilized to a substrate. For example, the same substrate can comprise allele-specific oligonucleotide probes for detecting at least 10, 100 or all of the polymorphisms shown in Tables 2-11. Optional additional components of the kit include, for example, restriction enzymes, reverse-transcriptase or polymerase, the substrate nucleoside triphosphates, means used to label (for example, an avidin-enzyme conjugate and enzyme substrate and chromogen if the label is biotin), and the appropriate buffers for reverse transcription, PCR, or hybridization reactions. Usually, the kit also contains instructions for carrying out the methods.

EXAMPLES

The polymorphisms set forth in this application were identified by hybridization to tiling arrays. Tiling arrays are described in PCT/US94/12305 (incorporated by reference in its entirety for all purposes). Tiling generally means the synthesis of a defined set of oligonucleotide probes that is

made up of a sequence complementary to the sequence to be analyzed (the "target sequence"), as well as preselected variations of that sequence. The variations usually include substitution at one or more base positions with one or more nucleotides. Tiling strategies are discussed in WO 95/11995 (incorporated by reference in its entirety for all purposes). With a tiled array containing 4L probes one can query every position in a nucleotide containing L number of bases. A 4L tiled array, for example, contains L number of sets of 4 probes, i.e. 4L probes. Each set of 4 probes contains the perfect complement to a portion of the target sequence with a single substitution for each nucleotide at the same position in the probe. See also Chee et. al., *Science* October, 1996.

To detect the novel sequence tagged polymorphic sites provided in this application, we designed a $P^{25,13}$ (25-mer probes having the interrogation position at base 13) 4L tiling array for the G6PD locus. Because the G6PD locus contains a large number of Alu sequences (repeat sequences), we simplified the tiled probe array by not probing the repetitive Alu sequences. To generate target sequence fragments, blood was collected from 10 individuals. Long range PCR amplification was carried out on genomic DNA. The amplicons were labeled, fragmented, and used to determine hybridization to the array.

Table 1

	1	2	3	4	5	6	7	8	9	10
M1	G	G	C	T	C	T	G	C	G	G
M2	A	A	C	T	C	A	T	C	G	G
M3	A	A	C	T	C	A	T	C	G	G
M4	A	A	C	G	C	A	G	C	G	A
M5	G	G	G	T	C	A	G	C	G	G
M6	A	A	C	T	T	A	G	C	A	G
M7	G	G	C	T	C	A	G	C	G	G
M8	G	G	C	T	C	A	G	C	G	G
M9	A	A	C	T	T	A	G	C	A	G
M10	G	G	C	T	C	A	G	T	G	G

Table 2

Starting Sequence	T	G	A	G	C	A	A	C	A	G	T	G	G	A	A	A	T	T	T	G
M1	T	G	A	G	C	A	A	C	A	G	T	G	G	A	A	A	T	T	T	G
M10	T	G	A	G	C	A	A	C	A	G	T	G	G	A	A	A	T	T	T	G
M2	T	G	A	G	C	A	A	C	A	A	T	G	G	A	A	A	T	T	T	G
M3	T	G	A	G	C	A	A	C	A	A	T	G	G	A	A	A	T	T	T	G
M4	T	G	A	G	C	A	A	C	A	A	T	G	G	A	A	A	T	T	T	G
M5	T	G	A	G	C	A	A	C	A	G	T	G	G	A	A	A	T	T	T	G
M6	T	G	A	G	C	A	A	C	A	A	T	G	G	A	A	A	T	T	T	G
M7	T	G	A	G	C	A	A	C	A	G	T	G	G	A	A	A	T	T	T	G
M8	T	G	A	G	C	A	A	C	A	G	T	G	G	A	A	A	T	T	T	G
M9	T	G	A	G	C	A	A	C	A	A	T	G	G	A	A	A	T	T	T	G

Table 3

Starting Sequence	G	C	A	G	T	T	T	G	A	**G**	T	G	T	C	T	C	T	G	G	T
M1	G	C	A	G	T	T	T	G	A	**G**	T	G	T	C	T	C	T	G	G	T
M10	G	C	A	G	T	T	T	G	A	**G**	T	G	T	C	T	C	T	G	G	T
M2	G	C	A	G	T	T	T	G	A	**A**	T	G	T	C	T	C	T	G	G	T
M3	G	C	A	G	T	T	T	G	A	**A**	T	G	T	C	T	C	T	G	G	T
M4	G	C	A	G	T	T	T	G	A	**A**	T	G	T	C	T	C	T	G	G	T
M5	G	C	A	G	T	T	T	G	A	**G**	T	G	T	C	T	C	T	G	G	T
M6	G	C	A	G	T	T	T	G	A	**A**	T	G	T	C	T	C	T	G	G	T
M7	G	C	A	G	T	T	T	G	A	**G**	T	G	T	C	T	C	T	G	G	T
M8	G	C	A	G	T	T	T	G	A	**G**	T	G	T	C	T	C	T	G	G	T
M9	G	C	A	G	T	T	T	G	A	**A**	T	G	T	C	T	C	T	G	G	T

Table 4

Starting Sequence	G	T	A	A	A	T	G	C	T	**C**	T	G	C	A	A	A	T	A	A	C
M1	G	T	A	A	A	T	G	C	T	**C**	T	G	C	A	A	A	T	A	A	C
M10	G	T	A	A	A	T	G	C	T	**C**	T	G	C	A	A	A	T	A	A	C
M2	G	T	A	A	A	T	G	C	T	**C**	T	G	C	A	A	A	T	A	A	C
M3	G	T	A	A	A	T	G	C	T	**C**	T	G	C	A	A	A	T	A	A	C
M4	G	T	A	A	A	T	G	C	T	**C**	T	G	C	A	A	A	T	A	A	C
M5	G	T	A	A	A	T	G	C	T	**G**	T	G	C	A	A	A	T	A	A	C
M6	G	T	A	A	A	T	G	C	T	**C**	T	G	C	A	A	A	T	A	A	C
M7	G	T	A	A	A	T	G	C	T	**C**	T	G	C	A	A	A	T	A	A	C
M8	G	T	A	A	A	T	G	C	T	**C**	T	G	C	A	A	A	T	A	A	C
M9	G	T	A	A	A	T	G	C	T	**C**	T	G	C	A	A	A	T	A	A	C

Table 5

Starting Sequence	G	G	C	T	C	C	A	A	G	**C**	G	G	T	G	C	C	C	G	G	C
M1	G	G	C	T	C	C	A	A	G	**C**	G	G	T	G	C	C	C	G	G	C
M10	G	G	C	T	C	C	A	A	G	**C**	G	G	T	G	C	C	C	G	G	C
M2	G	G	C	T	C	C	A	A	G	**C**	G	G	T	G	C	C	C	G	G	C
M3	G	G	C	T	C	C	A	A	G	**C**	G	G	T	G	C	C	C	G	G	C
M4	G	G	C	T	C	C	A	A	G	**C**	G	G	T	G	C	C	C	G	G	C
M5	G	G	C	T	C	C	A	A	G	**C**	G	G	T	G	C	C	C	G	G	C
M6	G	G	C	T	C	C	A	A	G	**T**	G	G	T	G	C	C	C	G	G	C
M7	G	G	C	T	C	C	A	A	G	**C**	G	G	T	G	C	C	C	G	G	C
M8	G	G	C	T	C	C	A	A	G	**C**	G	G	T	G	C	C	C	G	G	C
M9	G	G	C	T	C	C	A	A	G	**T**	G	G	T	G	C	C	C	G	G	C

Table 6

Starting Sequence	G	A	C	C	T	C	T	T	T	**A**	G	C	T	C	G	T	T	A	T	T
M1	G	A	C	C	T	C	T	T	T	**T**	G	C	T	C	G	T	T	A	T	T
M10	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T
M2	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T
M3	G	A	C	C	T	C	T	T	T	**A**	G	C	T	C	G	T	T	A	T	T
M4	G	A	C	C	T	C	T	T	T	**A**	G	C	T	C	G	T	T	A	T	T
M5	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T
M6	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T
M7	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T
M8	G	A	C	C	T	C	T	T	T	**A**	G	C	T	C	G	T	T	A	T	T
M9	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T

Table 7

Starting Sequence	G	G	G	C	C	T	C	A	A	**G**	A	T	T	T	G	A	T	T	T	C
M1	G	G	G	C	C	T	C	A	A	**G**	A	T	T	T	G	A	T	T	T	C
M10	G	G	G	C	C	T	C	A	A	**G**	A	T	T	T	G	A	T	T	T	C
M2	G	G	G	C	C	T	C	A	A	**T**	A	T	T	T	G	A	T	T	T	C
M3	G	G	G	C	C	T	C	A	A	**T**	A	T	T	T	G	A	T	T	T	C
M4	G	G	G	C	C	T	C	A	A	**G**	A	T	T	T	G	A	T	T	T	C
M5	G	G	G	C	C	T	C	A	A	**G**	A	T	T	T	G	A	T	T	T	C
M6	G	G	G	C	C	T	C	A	A	**G**	A	T	T	T	G	A	T	T	T	C
M7	G	G	G	C	C	T	C	A	A	**G**	A	T	T	T	G	A	T	T	T	C
M8	G	G	G	C	C	T	C	A	A	**G**	A	T	T	T	G	A	T	T	T	C
M9	G	G	G	C	C	T	C	A	A	**G**	A	T	T	T	G	A	T	T	T	C

Table 8

Starting Sequence	A	G	G	G	G	G	G	C	T	**T**	T	T	T	C	C	A	G	C	T	C
M1	A	G	G	G	G	G	G	C	T	**C**	T	T	T	C	C	A	G	C	T	C
M10	A	G	G	G	G	G	G	C	T	**T**	T	T	T	C	C	A	G	C	T	C
M2	A	G	G	G	G	G	G	C	T	**C**	T	T	T	C	C	A	G	C	T	C
M3	A	G	G	G	G	G	G	C	T	**C**	T	T	T	C	C	A	G	C	T	C
M4	A	G	G	G	G	G	G	C	T	**C**	T	T	T	C	C	A	G	C	T	C
M5	A	G	G	G	G	G	G	C	T	**C**	T	T	T	C	C	A	G	C	T	C
M6	A	G	G	G	G	G	G	C	T	**C**	T	T	T	C	C	A	G	C	T	C
M7	A	G	G	G	G	G	G	C	T	**C**	T	T	T	C	C	A	G	C	T	C
M8	A	G	G	G	G	G	G	C	T	**C**	T	T	T	C	C	A	G	C	T	C
M9	A	G	G	G	G	G	G	C	T	**C**	T	T	T	C	C	A	G	C	T	C

Table 9

Starting Sequence	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A
M1	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A
M10	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A
M2	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A
M3	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A
M4	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A
M5	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A
M6	G	C	C	T	C	C	T	T	C	A	T	T	C	T	A	C	G	A	C	A
M7	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A
M8	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A
M9	G	C	C	T	C	C	T	T	C	A	T	T	C	T	A	C	G	A	C	A

Table 10

Starting Sequence	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G
M1	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G
M10	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G
M2	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G
M3	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G
M4	A	G	G	G	T	G	C	G	C	A	T	C	C	T	C	A	C	C	T	G
M5	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G
M6	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G
M7	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G
M8	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G
M9	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G

Table 11

Starting Sequence	A	A	C	C	A	G	A	A	T	**T**	T	A	T	T	T	T	G	A	G	G
M1	A	A	C	C	A	G	A	A	T	**T**	T	A	T	T	T	T	G	A	G	G
M10	A	A	C	C	A	G	A	A	T	**T**	T	A	T	T	T	T	G	A	G	G
M2	A	A	C	C	A	G	A	A	T	**T**	T	A	T	T	T	T	G	A	G	G
M3	A	A	C	C	A	G	A	A	T	**T**	T	A	T	T	T	T	G	A	G	G
M4	A	A	C	C	A	G	A	A	T	**G**	T	A	T	T	T	T	G	A	G	G
M5	A	A	C	C	A	G	A	A	T	**T**	T	A	T	T	T	T	G	A	G	G
M6	A	A	C	C	A	G	A	A	T	**T**	T	A	T	T	T	T	G	A	G	G
M7	A	A	C	C	A	G	A	A	T	**T**	T	A	T	T	T	T	G	A	G	G
M8	A	A	C	C	A	G	A	A	T	**T**	T	A	T	T	T	T	G	A	G	G
M9	A	A	C	C	A	G	A	A	T	**T**	T	A	T	T	T	T	G	A	G	G

All publications and patent applications cited above are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication or patent application were specifically and individually indicated to be so incorporated by reference. Although the present invention has been described in some detail by way of illustration and example for purposes of clarity and understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims.

WHAT IS CLAIMED IS:

1 1 A nucleic acid segment of between 10 and 100 bases
2 of which at least 10 contiguous bases including a polymorphic
3 site are from any of the sequences shown in any of Tables 2-11,
4 or the complements thereof.

1 2. The nucleic acid segment of claim 1 that is DNA.

1 3. The nucleic acid segment of claim 1 that is RNA.

1 4 The segment of claim 1 that is less than 50 bases.

1 5. The segment of claim 1 that is less than 20 bases.

1 6. The segment of claim 1, wherein the sequence is a
2 sequence designated any of M1-M10 in any of Tables 2-11.

1 7. An allele-specific oligonucleotide that hybridizes
2 to a sequence shown in any of Tables 2-11 or its complement.

1 8. The allele-specific oligonucleotide of claim 7
2 that is a probe.

1 9. The allele-specific oligonucleotide of claim 8,
2 wherein a central position of the probe aligns with a
3 polymorphic site in the sequence.

1 10. The allele-specific oligonucleotide of claim 7 that
2 is a primer.

11. The allele-specific oligonucleotide of claim 10, wherein the 3' end of the primer aligns with a polymorphic site in the sequence.

12. An isolated nucleic acid comprising at least ten contiguous amino acids including the polymorphic site of an allelic variant of a starting sequence shown in any of Tables 2-11, or the complement thereof.

13. A method of analyzing a nucleic acid, comprising:
obtaining the nucleic acid from an individual; and
determining a base occupying a polymorphic site shown in any of Tables 2-11.

14. The method of claim 13, wherein the determining comprises determining a set of bases occupying a set of polymorphic sites shown in any of Tables 2-11.

15. The method of claim 14, wherein the nucleic acid is obtained from a plurality of individuals, and a base occupying one of the polymorphic sites is determined in each of the individuals, and the method further comprising testing each individual for the presence of a disease phenotype, and correlating the presence of the disease phenotype with the base.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US97/19665

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12Q 1/68; C07H 21/04
US CL :435/6; 536/23.5, 24.31, 24.33

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 91.1, 91.2, 183; 536/23.1, 23.5, 24.31, 24.33

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	SIGMA MOLECULAR BIOLOGY, 1989, page 54.	7-11
X	NIERMAN et al., Eds., ATCC/NIH Repository Catalogue of Human and Mouse DNA Probes and Libraries, Eighth Edition, 1994, pages 1-58.	1, 2, 12

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"B"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 02 JANUARY 1998	Date of mailing of the international search report 10 FEB 1998
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer BRADLEY L. SISSON Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*